HierFlat: flattened hierarchies for improving top-down hierarchical classification

Azad Naik¹ · Huzefa Rangwala¹

Received: 26 January 2017 / Accepted: 30 August 2017 / Published online: 12 September 2017
© Springer International Publishing AG 2017

Abstract Large-scale classification of structured data where classes are organized in a hierarchical structure is an important area of research. Top-down approaches that leverage the hierarchy during the learning and prediction phase are efficient for solving large-scale hierarchical classification. However, accuracy of top-down approaches is poor due to error propagation, i.e., prediction errors made at higher levels in the hierarchy cannot be corrected at lower levels. One of the main reasons behind errors at the higher levels is the presence of inconsistent nodes that are introduced due to the arbitrary process of creating these hierarchies by domain experts. In this paper, we propose two different data-driven approaches (local and global) for hierarchical structure modification that identifies and flattens inconsistent nodes present in the hierarchy. Our extensive empirical evaluation of the proposed approaches on several image and text datasets with varying distribution of features, classes and training instances per class shows improved classification performance over competing hierarchical modification approaches. Specifically, we see an improvement up to 7% in Macro-F1 score with our best approach over best top-down baseline. We also show that our modified hierarchy leads to improved classification performance for classes with few training samples in comparison with flat approach. Further, our experimental evaluation shows that combining the prediction from flat and

hierarchical model in an ensemble setting results in better classification performance.

Keywords Hierarchical classification · Inconsistency · Error propagation · Flattening · Logistic regression · Support vector machine · Hybrid prediction

1 Introduction

Hierarchies (taxonomies) are the most common and useful data structures for organizing large volume of datasets in various application domains including bioinformatics¹ for organizing gene sequences, ImageNet² for organizing images and Yahoo!³ web directories for organizing web documents. Given the hierarchy of classes, the hierarchical classification (HC) problem deals with learning models by utilizing (or ignoring) the hierarchical structure to automatically classify unlabeled test instances (examples) into relevant classes (categories). Over the years, HC has gained immense interest among researchers and is evident from the various large-scale online prediction challenges such as LSHTC⁴ (web-page documents classification), BioASQ⁵ (published medical abstracts classification) and ILSVRC⁶ (large-scale image classification).

In the past, various methods have been developed to improve the HC performance [2]. One of the simplest meth-

This paper is an extended version of the DSAA'2016 Long Presentation Paper, "Inconsistent Node Flattening for Improving Top-Down Hierarchical Classification" [1].

✉ Azad Naik
anaik3@gmu.edu

Huzefa Rangwala
rangwala@cs.gmu.edu

¹ George Mason University, Fairfax, VA, USA

¹ <http://geneontology.org/>.

² <http://image-net.org/>.

³ <http://dir.yahoo.com/>.

⁴ <http://lshtc.iit.demokritos.gr/>.

⁵ <http://www.bioasq.org/>.

⁶ <http://image-net.org/challenges/LSVRC/>.

ods is to learn binary one-vs-rest classifier for each of the leaf categories, ignoring the hierarchical relationships. This method is referred as flat classification. Other methods involve use of the hierarchies (see Sect. 2) during the learning and prediction process. Hierarchies provide useful structural relationships (such as parent–child and siblings) among different classes that can be exploited for learning a general classification models. Previously, researchers have demonstrated the usefulness of hierarchies for classification and have obtained promising results [3–12]. However, in many situations hierarchies used for learning models are not consistent due to the presence of *inconsistent* nodes resulting in excessive error propagation. As such, HC approaches are outperformed by the flat classifiers that completely ignore the hierarchy [13, 14].

Flat classifiers, though effective in some cases, suffer from two major issues: (i) During the prediction phase, flat classifiers invoke all the models for leaf categories and are considerably slower than top-down HC approaches, in which only the models in the relevant path are invoked. (ii) For large-scale HC problems, it is challenging to learn classification models that can discriminate between large number of classes. This problem is worse for datasets with skewed class distributions where plenty of classes have very few examples for training (rare categories problem) [15]. Large-scale datasets show power-law distribution of examples per category [16]. Considering these issues, the focus of this paper is to improve top-down HC approaches, which are computationally feasible for large-scale datasets and handle the imbalance problem by utilizing structural relationships information.

The main drawback of top-down HC approaches that contributes to their inferior classification performance is error propagation—compounding of errors from misclassifications at higher levels which cannot be rectified at the lower levels. This problem can be alleviated to certain extent by restructuring (modifying) the hierarchy to remove set of inconsistent nodes that causes performance deterioration. In this paper, our main contribution includes development of data-driven approaches for removing inconsistencies in the expert-defined (original) hierarchy leading to a hierarchy that achieves higher classification performance irrespective of the HC approach used for training. We propose a *flattening* approach where inconsistent nodes are selectively removed from the hierarchy. The criterion for flattening a node is based on the optimal regularized risk minimization objective value attained by the model trained for that node on a separate validation set. If the objective value for a node n is above a certain threshold, then we flatten n , i.e., we remove n from the hierarchy and add its children to n 's parent node. Based on the strategy adapted for identifying inconsistent nodes, we propose two different approaches for inconsistent nodes flattening (INF) from the hierarchy: (i) local approach (Level-

INF) that computes a level-wise cutoff threshold for each level in the hierarchy and (ii) global approach (Global-INF) that computes a global threshold for the entire hierarchy.

Experimental comparisons of top-down HC approach on our proposed modified hierarchy show statistically significant performance improvement in comparison with the baseline hierarchy (expert-defined) and other competitive methods for hierarchy modification [17, 18]. We also performed detailed analysis to show that the reduction in misclassification error at higher levels with our proposed hierarchy modification approach leads to reduced error propagation and hence better classification performance. In comparison with flat classification, our approach is more accurate for classes with few training instances per class (rare categories) and is computationally efficient for large-scale HC problems during the prediction phase.

To summarize, the main contributions of our work are as follows:

1. We propose a *local* and *global* methods for flattening inconsistent nodes in the hierarchy that gives comparatively better classification performance over baseline expert-defined hierarchy and other competing hierarchy modification approaches [17, 18] using probabilistic classifier such as logistic regression (LR) and large-margin classifier such as support vector machine (SVM).
2. We perform a case study and experiments on wide range of image and text datasets with varying characteristics to show the strength and effectiveness of our proposed modification approaches.
3. We perform level-wise error analysis to compare the error propagation with modified and the baseline hierarchy.
4. We analyze the performance improvement achieved with proposed modified hierarchy over the flat method on varying distribution of training examples per class.
5. We perform a study that combines the flat and top-down model prediction in an ensemble setting (hybrid prediction).

2 Literature review

There has been a large body of research focusing on the HC problem. Besides completely ignoring the hierarchy (i.e., flat classifiers), one class of HC methods solves various local subproblems that train individual classifier(s) for each of the nodes (or parent nodes) in the hierarchy or learn classifiers for each of the levels. These methods are referred as local classification because only local structural relationship information is used during training these classifiers. To predict the labels of instances, top-down local hierarchical methods proceed by selecting the most relevant node at the topmost level and then recursively selecting the best child node until

Fig. 1 Various hierarchical structures (b)–(f) obtained after flattening some of the nodes (or levels) from the original hierarchy shown in (a). ‘LN’ denotes the leaf node to which examples are assigned. Internal nodes in each level of the original hierarchy are represented with

different symbol and color for ease of understanding. **a** Expert-defined (original) hierarchy. **b** Flat hierarchy. **c** Top-level flattened (TLF) hierarchy. **d** Bottom-level flattened (BLF) hierarchy. **e** Multiple-level flattened (MLF) hierarchy. **f** Inconsistent node flattened (INF) hierarchy

a leaf category is reached, which is the final predicted label. Local approaches are more popular due to their computational benefits [2,6]. Contrary to local classification, global classification methods [3,19] learn a single complex model over all the nodes in the hierarchy and are computationally more expensive than flat and local methods. Therefore, in this paper we focus on top-down local classification methods for training models and predicting labels.

Some of the earlier studies focus on exploiting hierarchies among categories for the purpose of classification [5,20–22],

but the number of categories was limited to a few hundreds. One of the earlier breakthroughs in the field of hierarchical text categorization was by Koller et al. [6]. This approach used a divide and conquer paradigm for solving the HC problem which can easily be adapted in large-scale settings. Following this, numerous approaches have been developed to improve HC performance on a large datasets, both accuracy and scalability. Liu et al. [23] studied the classification performance on Yahoo! web taxonomy using a SVM-based method that scales for millions of categories. Gopal et al. [3]

used a regularization term within the optimization function to capture the parent–child relationships in the hierarchy. This approach referred as HR-LR and HR-SVM shows improved classification performance, but the training procedure for large-scale problem requires a distributed implementation and map-reduce supported infrastructure. Xue et al. [12] proposed a two-stage approach. For each test document, in the first stage a set of candidate categories are retrieved based on similarity to the test document. Then the second stage builds a classifier on the hierarchy restricted to the set of categories fetched in first stage and classifies the test document using the restricted hierarchy. Although pruning reduces the hierarchy to a manageable size, one severe drawback of this approach is having to train a different classifier for each test document which is expensive. Other works in the field of HC can be found in a detailed survey by Silla et al. [2].

2.1 Hierarchy modification

Most HC approaches rely on hierarchical relationships for learning complex models to improve the classification performance. However, the performance can be negatively impacted if the hierarchy used during learning models is inconsistent. Therefore, it is of utmost importance to generate an improved hierarchical representation that is suitable for classification prior to learning models. Inconsistencies in the hierarchy are due to the following reasons:

1. Hierarchies are designed for efficient search and navigation without considering HC performance.
2. Hierarchical groupings of categories are done based on semantics, whereas classification depends on data characteristics such as *term-frequency*.
3. Multiple hierarchies are possible for the same dataset (such as SCOP and CATH [24] hierarchy for protein structures). However, there is no consensus regarding which hierarchy is better for classification.
4. Consistent hierarchy design for dataset with large number of categories is prone to errors.

Several approaches that restructure the hierarchy have been developed in past. Level flattening [17] is one of the approaches used in earlier works of hierarchy modification, where some of the levels are flattened (removed) from the original hierarchy prior to learning models. Based on levels that are flattened, various methods of modification exist. Top-level flattening (TLF) as shown in Fig. 1c modifies the hierarchy by flattening the top level in the original hierarchy. Model learning and prediction for flattened hierarchy is done in similar fashion as top-down methods. Bottom-level flattening (BLF) and multiple-level flattening (MLF), shown in Fig. 1d, e are similar methods of hierarchy modification where bottom and multiple levels are removed, respectively.

As done in Wang et al. [17], we removed the first and third levels for evaluating the MLF approach.

Babbar et al. [18] proposed a maximum-margin-based strategy for hierarchy modification. This method selectively removes some of the inconsistent nodes in the hierarchy based on margins rather than removing complete levels. Hierarchy modification using this approach, shown in Fig. 1f, is beneficial for classification and has been theoretically justified [25]. We followed a similar approach for hierarchy modification. However, our method differs in following two aspects: (i) We developed a more systematic and elegant approach for threshold selection to identify the inconsistent nodes for flattening. Our approach is based on deviation from mean that is empirically tuned for each dataset, and (ii) we also considered a global perspective of the hierarchy (Global-INF) for threshold selection to identify the inconsistent set of nodes. This approach is more intuitive and realistic measure for threshold selection because it prevents excessive flattening of the nodes that is just based on local decisions, thereby allowing the benefits of leveraging the hierarchy during the model learning and classification process, especially for rare categories (see Sect. 6 for justification).

Hierarchy modification using a supervised learning approaches is also proposed in the literature [15,21], where the hierarchy is gradually modified to achieve better hierarchy for improving the classification performance. These methods have an expensive evaluation cost that needs to be performed after each modification, making it computationally in-feasible for large-scale settings. Hence, we do not compare our approach to these methods. Other competitive methods that involve restructuring the hierarchy are developed by us and appear in an arXiv publication [26].

3 Methods

Table 1 summarizes the common notations used in this paper. We use bold letters to indicate vector variables.

3.1 Hierarchical classification

Given, a hierarchy $\mathcal{H}$ we train a binary one-vs-rest classifiers for each of the node $n \in \mathcal{N}$ —to discriminate its positive examples from the examples of other nodes (i.e., negative examples) in the hierarchy. We followed the ‘inclusive policy’ for training classifiers, where all the descendant categories of node n (including node itself) are considered as positive examples and the remaining categories as negative examples [27]. In this paper, we explore two choices for training the model, namely logistic regression (LR) and support vector machine (SVM) [3]. The LR model uses logistic loss to minimize the empirical risk, whereas SVM model uses hinge loss. Both these models are coupled with l_2 -norm

Table 1 Notation description

Symbol	Description
$\mathbb{R}$	Set of real numbers
d	Dimensionality of input vector
$\mathcal{H}$	Expert-defined (original) hierarchy
N	Total number of training examples
$\mathcal{L}$	Set of leaf categories
$\mathbf{x}_i \in \mathbb{R}^d$	Input vector for i -th training example
$\hat{\mathbf{x}}_i \in \mathbb{R}^d$	i -th test example
$y_i \in \mathcal{L}$	True label for i -th training example
$\hat{y}_i \in \mathcal{L}$	Predicted label for i -th test example
$y_i^n \in \pm 1$	Binary label used for i -th training example to learn weight vector for n -th node in $\mathcal{H}$, $y_i^n = 1$ iff $y_i = n$, -1 otherwise
$\hat{y}_i^n \in \pm 1$	Predicted label for i -th test example at n -th node in the hierarchy, $\hat{y}_i^n = 1$ iff prediction is positive, -1 otherwise
Θ_n	Weight vector for n -th node in the hierarchy
$C > 0$	Misclassification penalty parameter
$\mathcal{C}(n)$	Set of children of n -th node in the hierarchy
$\mathcal{A}(n)$	Set of ancestors of n -th node (including root)
f_n^*	Optimal objective function value for n -th node obtained using validation dataset. We have dropped the subscript n at some places for ease of description
$\mathcal{H}_L$	Modified hierarchy after level-wise inconsistent node flattening
$\mathcal{H}_G$	Modified hierarchy after global inconsistent node flattening
$\mathcal{N}_k$	Set of nodes at k -th level in $\mathcal{H}$
$\mathcal{N}$	Set of nodes (except root) in $\mathcal{H}$
$\mathcal{I}_L$	Set of inconsistent nodes identified using level-wise flattening method
$\mathcal{I}_G$	Set of inconsistent nodes identified using global flattening method
$\mu(\mathcal{S})$	Mean of samples in set $\mathcal{S}$
$\sigma(\mathcal{S})$	Standard deviation of samples in set $\mathcal{S}$
$\mathcal{S}_k$	Set of f^* values for node at k -th level in $\mathcal{H}$
$\mathcal{S}$	Set of f^* values for all nodes (except root) in $\mathcal{H}$
τ_k	Threshold limit for identifying inconsistent nodes at k -th level in $\mathcal{H}$
τ	Global threshold limit for identifying inconsistent nodes in $\mathcal{H}$
$\psi_k \geq 0$	Fitness parameter for level-wise threshold selection at k -th level in $\mathcal{H}$
$\psi \geq 0$	Fitness parameter for global threshold selection in $\mathcal{H}$

term (denoted by $\|\cdot\|_2^2$) to control the model complexity and prevent from overfitting. The objective function f_n for training a model corresponding to node n using LR and SVM objective is provided in Eqs. (1) and (2), respectively.

$$f_n = \min_{\Theta_n} \left[C \sum_{i=1}^N \log \left(1 + \exp \left(-y_i^n \Theta_n^T \mathbf{x}_i \right) \right) + \frac{1}{2} \|\Theta_n\|_2^2 \right] \quad (1)$$

$$f_n = \min_{\Theta_n} \left[C \sum_{i=1}^N \left| 1 - y_i^n \Theta_n^T \hat{\mathbf{x}}_i \right|_+ + \frac{1}{2} \|\Theta_n\|_2^2 \right] \quad (2)$$

where $|a|_+$ denotes the hinge loss given by $\max(0, a)$.

For each node n , we solve Eq. (1) [or Eq. (2)] to obtain the optimal weight vector denoted by Θ_n . The complete set of parameters for all the nodes $\{\Theta_n\}_{n \in \mathcal{N}}$ constitutes the learned model for the hierarchical top-down classifier.

For a test example with feature vector $\hat{\mathbf{x}}_i$, the top-down classifier predicts the class label $\hat{y}_i \in \mathcal{L}$ as shown in Eq. (3). Essentially, the algorithm starts at the root node and recursively selects the best child nodes till it reaches a terminal node belonging to the set of leaf categories $\mathcal{L}$.

$$\hat{y}_i = \left\{ \begin{array}{l} \textbf{initialize } p := \text{root} \\ \textbf{while } p \notin \mathcal{L} \\ \quad p := \textbf{argmax}_{q \in \mathcal{C}(p)} \Theta_q^T \hat{\mathbf{x}}_i \\ \textbf{return } p \end{array} \right\} \quad (3)$$

Although in this paper we have explored two of the most widely used classifier (LR and SVM), other binary or multi-class classifiers [28–30] can also be used to perform HC.

3.2 Inconsistent node flattening

Motivation: Gao et al. [25] showed theoretically that for any classifier that correctly classifies m random input-output pairs using a set of $\mathcal{D}$ decision nodes with margins $\{\gamma_n, \forall n \in \mathcal{D}\}$, the generalization error with probability estimates greater than $(1 - \zeta)$ is bounded by the expression shown in Eq. (4).

$$\frac{130R^2}{m} \left[\sum_{n \in \mathcal{D}} \left(\frac{1}{\gamma_n^2} \right) \log(4em) \log(4m) + |\mathcal{D}| \log(2m) - \log \left(\frac{2}{\zeta} \right) \right] \quad (4)$$

where R is the radius of the ball containing the distribution's support.

This bound provides two significant strategies in designing our approach to reduce the generalization error: (i) Increase the margin γ_n for learned models at node $n \in \mathcal{N}$ in the hierarchy, or (ii) decrease the number of decision nodes $|\mathcal{D}|$ involved in making the prediction. For achieving the optimum classification performance, we need to balance the trade-off between the margin γ_n and the number of decision nodes $|\mathcal{D}|$. Two of the extreme cases for learning hierarchical

classifiers are flat and top-down methods. For flat classifiers, we have to make single decision (i.e., $|\mathcal{D}| = 1$) but margin width γ_n is presumably small due to the large number of leaf categories that needs to be distinguished, which makes it difficult to obtain large margin. For top-down hierarchical classifiers, we have to make a series of decisions from root to leaf nodes (i.e., $|\mathcal{D}| \geq 1$) but margin γ_n is larger due to the fewer number of categories that needs to be distinguished at each of the decision nodes. Motivated by this trade-off, in this work we propose a method that removes some of the inconsistent nodes in the hierarchy $\mathcal{H}$, and thereby, increasing the value of margin γ_n for learned models at node n in the hierarchy, while minimizing the number of decision nodes to classify an unlabeled test instances.

In order to improve the effectiveness of classification, we need to identify these inconsistent nodes and flatten them. We mark a node n within the hierarchy as inconsistent if the value of the objective function f_n^* becomes greater than a chosen threshold value. To get a more reliable estimate of the f_n^* , we first train the regularized LR models on a training set locally for each node and then compute the objective function on a separate validation set, which is different from the training set. The objective value on validation set for node n is denoted by f_n^* . We develop the following approaches for setting the threshold for flattening.

Level-wise inconsistent node flattened: In this approach, referred as Level-INF, we select a different threshold τ_k locally for each level k in the hierarchy. Algorithm 1 presents the level-wise approach that selects inconsistent nodes at each level in a top-down manner. The threshold τ_k for level k is computed as the sum of mean and ψ_k times the standard deviation of the set of values $\{f_n^*\}_{n \in \mathcal{N}_k}$, where ψ_k is a fitness parameter at level k that is empirically estimated for each dataset based on $\{f_n^*\}_{n \in \mathcal{N}_k}$ values (see Sect. 6.3) and $\mathcal{N}_k$ represents the set of nodes in level k . All nodes $n \in \mathcal{N}_k$ that satisfy $f_n^* > \tau_k$ are marked as inconsistent and added to the set of inconsistent nodes denoted by $\mathcal{I}_L$. This procedure is repeated for all levels of the hierarchy. Finally, we flatten the nodes in set $\mathcal{I}_L$ —remove $n \in \mathcal{I}_L$ and corresponding edges, and add edges from children of n to n 's parent node. The modified hierarchy thus obtained is denoted by $\mathcal{H}_L$. Using the modified hierarchy, we retrain a top-down classifier.

Global inconsistent node flattened: Different from Level-INF approach, which sets different thresholds for each level, the global method shown in Algorithm 2 computes a single threshold value for all levels. The threshold τ is computed as the sum of mean and ψ times the standard deviation of the set of value $\{f_n^*\}_{n \in \mathcal{N}}$, where ψ is a fitness parameter that is empirically estimated for dataset considering all $\mathcal{N}$ nodes f_n^* values. τ is used to identify the set of inconsistent nodes $\mathcal{I}_G$ in the hierarchy (i.e., all nodes n with $f_n^* > \tau$). The hierarchy obtained by flattening the nodes present in $\mathcal{I}_G$ is denoted by $\mathcal{H}_G$. Using the modified hierarchy, we retrain

Algorithm 1 Level-wise Inconsistent Node Flattened

Data: Original Hierarchy $\mathcal{H}$, input-output (x_i, y_i)
Result: Modified Hierarchy $\mathcal{H}_L$
 Train l_2 -regularized LR or SVM model in a top-down order
 /* Set of inconsistent node, initially empty */
 $\mathcal{I}_L := \Phi$;
for $k := 1 \dots \text{end_level}$ **do**
 /* Set of all nodes f_n^* values in the level */
 $S_k := \Phi$;
 for $n \in \mathcal{N}_k$ **do**
 $S_k := S_k \cup \{f_n^*\}$
 end
 $\tau_k := \mu(S_k) + \psi_k \sigma(S_k)$;
 /* Identify inconsistent node in level */
 for $n \in \mathcal{N}_k$ **do**
 if $(f_n^* > \tau_k \ \& \ n \notin \mathcal{L})$ **then**
 $\mathcal{I}_L := \mathcal{I}_L \cup \{n\}$
 end
 end
end
 /* New hierarchy with inconsistent nodes removed */
 $\mathcal{H}_L = \mathcal{H} - \{\mathcal{I}_L\}$;
return $\mathcal{H}_L$

a top-down classifier. In this paper, we refer to this approach as Global-INF.

Algorithm 2 Global Inconsistent Node Flattened

Data: Original Hierarchy $\mathcal{H}$, input-output (x_i, y_i)
Result: Modified Hierarchy $\mathcal{H}_G$
 Train l_2 -regularized LR or SVM model in a top-down order
 /* Set of inconsistent node, initially empty */
 $\mathcal{I}_G := \Phi$;
 /* Set of all nodes (except root) f_n^* values in $\mathcal{H}$ */
 $S := \Phi$;
for $n \in \mathcal{N}$ **do**
 $S := S \cup \{f_n^*\}$
end
 $\tau := \mu(S) + \psi \sigma(S)$;
 /* Identify inconsistent node in $\mathcal{H}$ */
for $n \in \mathcal{N}$ **do**
 if $(f_n^* > \tau \ \& \ n \notin \mathcal{L})$ **then**
 $\mathcal{I}_G := \mathcal{I}_G \cup \{n\}$
 end
end
 /* New hierarchy with inconsistent nodes removed */
 $\mathcal{H}_G = \mathcal{H} - \{\mathcal{I}_G\}$;
return $\mathcal{H}_G$

3.3 Hybrid prediction

One of the major issues faced by top-down approaches is that the prediction errors made at the higher levels cannot be rectified at the lower levels. In order for the top-down prediction algorithm [given in Eq. (3)] to predict the correct label for a test example and avoid error propagation, all the classifiers in its path must classify correctly. The effect of this error propagation worsens with the increase in depth of the hierar-

Table 2 Dataset statistics

Dataset	Domain	#Total node	#Leaf node	#Levels	#Training	#Testing	#Features
CLEF	Image	88	63	4	10,000	1,006	80
DIATOMS	Image	399	311	4	1,940	993	371
IPC	Text	553	451	4	46,324	28,926	1,123,497
DMOZ-SMALL	Text	2,388	1,139	6	6,323	1,858	51,033
DMOZ-2010	Text	17,222	12,294	6	128,710	34,880	381,580
DMOZ-2012	Text	13,963	11,947	6	383,408	103,435	348,548

chy, leading to poor performance of top-down models [15]. To overcome this problem, we propose a hybrid approach for prediction that combines the prediction made by flat and top-down classifiers. Essentially, in this method we consider the predictions obtained from both the flat and top-down classifiers and in case of disagreement in label assignment, we choose the most confident prediction based on probability estimates. In our experimental evaluations, the hybrid approach performed better than both the flat and top-down methods.

Let $(P_i^F, \hat{y}_i^F)$ denote the probability score and predicted label pair obtained from flat classifier for i -th test instance, and $(P_i^{TD}, \hat{y}_i^{TD})$ denotes the corresponding values for top-down classifier. Since label for top-down model is obtained by firing multiple classifiers in the top-down path, we compute the final probability score P_i^{TD} as the multiplication of probabilities predicted by the classifiers in the path [31], as shown in Eq. (5).

$$P_i^{TD} = \prod_{k \in \mathcal{A}(\hat{y}_i^{TD})} p_i^k \quad (5)$$

where p_i^k is the probability score obtained for i -th instance at k -th level for the selected top-down path and is computed as shown in Eq. (6).

$$p_i^k = \frac{\exp(\Theta_k^T \hat{\mathbf{x}}_i)}{\exp(\Theta_k^T \hat{\mathbf{x}}_i) + \sum_{n \in \xi(k)} \exp(\Theta_n^T \hat{\mathbf{x}}_i)} \quad (6)$$

Similarly, to compute the probability score for flat classifier P_i^F , we average the prediction over all labels as shown in Eq. (7).

$$P_i^F = \frac{\exp(\arg\max_{n \in \mathcal{L}} \Theta_n^T \hat{\mathbf{x}}_i)}{\sum_{n \in \mathcal{L}} \exp(\Theta_n^T \hat{\mathbf{x}}_i)} \quad (7)$$

In case $\hat{y}_i^F \neq \hat{y}_i^{TD}$, we obtain the final prediction by assigning the label with highest probability score as shown in Eq. (8).

$$\text{Final prediction}(\hat{y}_i) = \begin{cases} \hat{y}_i^F, & P_i^F > P_i^{TD} \\ \hat{y}_i^{TD}, & \text{otherwise} \end{cases} \quad (8)$$

4 Experimental evaluations

4.1 Datasets

We have used text and image datasets for evaluating the performance of our proposed approaches. Various statistics of the datasets used in our experiments are listed in Table 2. All these datasets are single-labeled, and the examples are mandatorily assigned to the leaf nodes in the hierarchy. For all text datasets, we have applied the tf-idf transformation with l_2 -norm normalization to the word-frequency feature vector. Description of the used datasets is as follows:

4.1.1 Image datasets

CLEF [32]—Medical images annotated with medical applications codes. Each image is represented by the 80 features that are extracted using local distribution of edges.

DIATOMS [33]—Diatom images that was created as the part of the ADIAC project. Features for each image are created using various feature extraction techniques mentioned in [33]. Further, we have preprocessed the original dataset by removing the examples that belongs to the internal nodes.

4.1.2 Text datasets

IPC⁷—Collection of patent documents organized in international patent classification hierarchy.

DMOZ-SMALL, DMOZ-2010 and DMOZ-2012⁸—Multiple web documents organized into various classes using the hierarchical structure. It is subset of the web pages from open directory project and has been released as the part of the LSHTC⁹ challenge. For evaluating the DMOZ-2010 and DMOZ-2012 datasets we have used the provided test

⁷ <http://www.wipo.int/classifications/ipc/en/>.

⁸ <http://dmoz.org>.

⁹ <http://lshtc.iit.demokritos.gr/>.

Fig. 2 Distribution of DMOZ datasets visualizing majority of the classes with rare categories. **a** DMOZ-SMALL. **b** DMOZ-2010. **c** DMOZ-2012

split and prediction scores are obtained using the web portal interface^{10,11} that was used during the competition. DMOZ datasets are highly imbalanced with more than 75% of the classes belonging to rare categories (i.e., classes having ≤ 10 examples) as shown in Fig. 2.

4.2 Evaluation metrics

For evaluating the performance of HC models, we have used different flat and hierarchy-based metrics used in the online prediction challenges such as LSHTC and BioASQ. Description about various metrics used in our evaluation is as follows:

Flat evaluation measures: Standard set-based metrics [34] Micro- F_1 (μF_1) and Macro- F_1 (MF_1) are used for evaluating the performance of various methods. To compute μF_1 , we sum up the category specific true positives (TP_c), false positives (FP_c) and false negatives (FN_c) for different categories and compute the score as:

$$\begin{aligned} P &= \frac{\sum_{c \in \mathcal{L}} TP_c}{\sum_{c \in \mathcal{L}} (TP_c + FP_c)}, \\ R &= \frac{\sum_{c \in \mathcal{L}} TP_c}{\sum_{c \in \mathcal{L}} (TP_c + FN_c)}, \\ \mu F_1 &= \frac{2PR}{P + R} \end{aligned} \quad (9)$$

Unlike μF_1 , MF_1 gives equal weight to all the categories so that the average score is not skewed in favor of the larger categories. It is defined as follows:

$$\begin{aligned} P_c &= \frac{TP_c}{TP_c + FP_c}, R_c = \frac{TP_c}{TP_c + FN_c} \\ MF_1 &= \frac{1}{|\mathcal{L}|} \sum_{c \in \mathcal{L}} \frac{2P_c R_c}{P_c + R_c} \end{aligned} \quad (10)$$

where $|\mathcal{L}|$ is the number of leaf categories.

Hierarchical evaluation measures: Different from flat measures that penalizes each of the misclassified examples

equally, hierarchical measures [35] take into consideration hierarchical distance between the true label and predicted label for evaluating the classifier performance. In general, misclassifications that are closer to the actual class are less severe than misclassifications that are farther from the true class with respect to the hierarchy (e.g., an example from hockey class misclassified as the baseball class is less severe in comparison with the hockey misclassified as cat). The hierarchy-based measures include hierarchical F_1 (hF_1) [harmonic mean of hierarchical precision (hP) and hierarchical recall (hR)] and tree-induced error (TE) [22], which are defined as follows:

$$\begin{aligned} hP &= \frac{\sum_{i=1}^N |\mathcal{A}(\hat{y}_i) \cap \mathcal{A}(y_i)|}{\sum_{i=1}^N |\mathcal{A}(\hat{y}_i)|}, \\ hR &= \frac{\sum_{i=1}^N |\mathcal{A}(\hat{y}_i) \cap \mathcal{A}(y_i)|}{\sum_{i=1}^N |\mathcal{A}(y_i)|}, \\ hF_1 &= \frac{2 * hP * hR}{hP + hR} \end{aligned} \quad (11)$$

$$TE = \frac{1}{N} \sum_{i=1}^N \delta(\hat{y}_i, y_i) \quad (12)$$

where $\mathcal{A}(\hat{y}_i)$ and $\mathcal{A}(y_i)$ are, respectively, the set of ancestors of the predicted and true labels which include the label itself, but do not include the root node. $\delta(a, b)$ gives the length of the undirected path between categories a and b in the graph. For TE lower scores are better, whereas for all other measures higher scores are better.

Note that for consistent evaluation of hierarchical measures we have used the original hierarchy for all methods unless where noted.

4.3 Experimental protocol

In all the experiments, we have divided the training dataset into train and small validation dataset in the ratio 90:10. Each experiment was run five times with different sets of train and validation split chosen randomly. Testing is done on an independent held-out dataset as provided by these benchmarks. The model is trained by choosing misclassification penalty

¹⁰ <http://lshtc.iit.demokritos.gr/node/81>.

¹¹ http://lshtc.iit.demokritos.gr/LSHTC3_oracleUpload.

parameter (C) in the set $[10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3]$. The best parameter is selected using a validation set. The best parameters are used to retrain the models on the entire training set, and the performance is measured on a held-out test set. For the INF methods, we compute and save the f_n^* value for each node in the hierarchy using a validation set. Setting the threshold as $\mu + \psi\sigma$ (or $\mu + \psi_k\sigma$ for k -th level in Level-INF approach), we remove the inconsistent nodes where best value of fitness parameter ψ (or ψ_k) is computed empirically for each dataset (see Sect. 6.3). All experiments were conducted using a modified version of liblinear¹² software [36] and were run on ARGO, a research computing cluster provided by the Office of Research Computing (URL: <http://orc.gmu.edu>), at George Mason University, VA. The source code, data description and experimental settings used in this paper are made available at our website.¹³

5 Comparative approaches

5.1 Top-down hierarchical methods

For all top-down hierarchical baselines, we train a binary one-vs-rest classifiers for each of the node (except root) in the hierarchy and predictions are made starting from the root node and recursively selecting the best scoring child nodes until a leaf node is reached [see Eq. (3)]. Depending upon the hierarchy that we use during the training and prediction process, we compare with the following baselines.

Top-down (TD): Original hierarchy provided by the domain experts is used for classifiers' training and label prediction. Based on the model used for training, we have TD-LR and TD-SVM where we train LR and SVM model, respectively.

Level flattening: Modified hierarchy obtained by flattening different level(s) is used instead of original hierarchy. Depending on level(s) flattened, we have top-level flattening (TLF), bottom-level flattening (BLF), multiple-level flattening (MLF) hierarchy as discussed in Sect. 2.1.

Maximum-margin-based taxonomy adaptation (MTA): Original hierarchy is modified using the margin value computed at each node in the hierarchy as described in Babbar et al. [18].

5.2 Flat methods

Logistic regression (LR): We train binary one-vs-rest regularized LR classifiers for each of the leaf categories, ignoring the hierarchical structure. The prediction decision $\hat{y}$ for unlabeled test instance $\mathbf{x}$ is based on the maximum prediction

score achieved when compared across the one-vs-rest classifiers as shown in Eq. (13).

$$\hat{y} = \underset{n \in \mathcal{L}}{\operatorname{argmax}} \Theta_n^T \mathbf{x} \quad (13)$$

Error-correcting output code (ECOC) [37]: This approach combines binary classifiers to exploit correlations and correct errors. Codewords are generated randomly with bits assigned for representing the hierarchical information between the categories. Experiments were done with codeword length varying from 32 to 1024 bits depending on the dataset. For testing an unlabeled example, the output codeword is compared to the codeword of each categories, and the one with the minimum hamming distance is selected as the class label for that example.

6 Results and discussion

6.1 Case study

To understand the quality of different hierarchical structures (expert-defined and flattened) for the newsgroup dataset shown in Fig. 3, we perform top-down HC using LR model with each of the hierarchy, separately. The dataset has 11,269 training instances, 7505 test instances and 20 classes. We evaluate each of the hierarchy by randomly selecting five different sets of training and test split in the same ratio as original dataset.

The results of classification performance are shown in Table 3. We can see that using these modified hierarchies substantially improves the classification performance in comparison with the baseline expert-defined hierarchy. On comparing the proposed approaches for flattening hierarchies, the classification performance obtained from using the Global-INF hierarchy is found to be significantly better than the Level-INF hierarchy. This is because Global-INF method has the flexibility of not removing any nodes (or removing all the nodes) from each of the level resulting in better set of inconsistent nodes and hence better classification performance. On contrary, Level-INF removes at least one node from each level which may not be optimal because all the nodes in the level might be useful for classification.

6.2 Comparison to top-down hierarchical baselines

Performance based on flat metrics: Table 4 (Table 5) shows the μF_1 and MF_1 performance comparison of our proposed hierarchical modification approaches with TD-LR (TD-SVM) and competitive hierarchy modification approaches. We have used LR and SVM as the baseline model for evaluation [3]. From table we can see that our proposed Global-INF

¹² <http://www.csie.ntu.edu.tw/~cjlin/liblinear/>.

¹³ <http://cs.gmu.edu/~mlbio/HierFlat>.

Fig. 3 Different hierarchical structures (b)–(c) obtained from expert-defined hierarchy (a). **a** Expert-defined hierarchy. **b** Level-INF modified hierarchy. **c** Global-INF modified hierarchy

approach consistently outperforms all other approaches for different datasets across all metrics. In fact, for the CLEF and DIATOMS image datasets we see a relative performance improvement of $>7\%$ in MF_1 scores on comparing Global-INF with the best top-down modification baseline, i.e., MTA. To validate the performance improvement, we conducted pairwise statistical significance tests between our best approach, Global-INF and best top-down baseline for all datasets except DMOZ-2010 and DMOZ-2012, where true test labels (and class-wise performance) are not available from the online evaluation. Specifically, we compute sign test for μF_1 [38] and nonparametric Wilcoxon rank test for MF_1 scores (it should be noted that significance tests are between two approaches for single run). In Fig. 4 we present the pairwise statistical comparisons of LR model for different approaches studied here on the DMOZ-SMALL dataset. The Global-INF consistently outperforms other baseline approaches studied here.

Overall, results of statistical tests show that Global-INF approach significantly outperforms the best baseline (and hence other baselines) for all the datasets (see Tables 4 and 5).

On comparing our two proposed approaches, Global-INF has better performance over Level-INF. This is because the Level-INF approach strictly enforces some of the nodes to be flattened from each levels although there f_n^* value may be much lower than the other nodes at different levels in the hierarchy and vice versa. In contrast, Global-INF approach takes all nodes f_n^* values into consideration while making a decision and hence it determines a better set of inconsistent nodes. MTA approach has poor performance due to the similar issues as with Level-INF approach. Performance of

Table 3 μF_1 , MF_1 and hF_1 performance comparison of different hierarchical structures on newsgroup dataset for LR model

Metric	TD-LR (Fig. 3a)	Proposed method	
		Level-INF (Fig. 3b)	Global-INF (Fig. 3c)
$\mu F_1(\uparrow)$	77.04 (0.18)	77.40 (0.25)	79.42 (0.12)
$MF_1(\uparrow)$	77.94 (0.04)	78.12 (0.16)	79.82 (0.07)
$hF_1(\uparrow)$	78.25 (0.13)	78.97 (0.14)	79.96 (0.09)

Table shows mean and (standard deviation) in bracket across five runs

level flattening approaches, viz. TLF, BLF and MLF, suffers because these methods remove the entire level(s) in the hierarchy and do not take into consideration whether any node in that level is important for HC. TD-LR approach has the worst performance because of the inconsistent nodes present in the original hierarchy that are negatively impacting the generalization capabilities of learned models at the higher levels (see Sect. 6.4), which results in error propagation.

Performance based on hierarchical metrics: Hierarchical evaluation metrics hF_1 and TE compute errors for misclassified examples based on the definition of a defined hierarchy. Table 6 (Table 7) shows the hF_1 and TE score for all top-down approaches evaluated over the original hierarchy and the modified hierarchy (obtained by flattening). We can see that our proposed approach, Global-INF outperforms other approaches because it is able to identify a better set of inconsistent nodes. On comparing the classification performance over the original hierarchy and the modified hierarchy, we can see that for most of the approaches classification on modified hierarchy show an improved performance. This is because

Table 4 μF_1 and MF_1 performance comparison of various top-down hierarchical baselines against our proposed approaches for LR model

Dataset	Metric	Hierarchical baselines					Proposed approaches	
		TD-LR	TLF	BLF	MLF	MTA	Level-INF	Global-INF
CLEF	μF_1 ($\uparrow$)	72.74 (0.43)	75.84 (0.32)	73.76 (0.32)	X	74.48 (0.42)	75.25 (0.55)	77.14 [†] (0.01)
	MF_1 ($\uparrow$)	35.92 (0.01)	38.45 (0.65)	40.93 (0.19)	X	39.53 (0.74)	39.89 (0.24)	46.54 [‡] (0.01)
DIATOMS	μF_1 ($\uparrow$)	53.27 (0.32)	56.93 (0.28)	53.27 (0.24)	X	58.36 (0.64)	58.32 (0.64)	61.31 [‡] (0.53)
	MF_1 ($\uparrow$)	44.46 (0.24)	45.17 (0.62)	44.30 (0.64)	X	45.21 (0.65)	48.77 (0.12)	51.85 [‡] (0.23)
IPC	μF_1 ($\uparrow$)	49.32 (0.32)	51.28 (0.61)	50.36 (0.64)	X	51.36 (0.32)	50.40 (0.32)	52.30 [†] (0.12)
	MF_1 ($\uparrow$)	42.51 (0.94)	44.99 (0.43)	43.74 (0.81)	X	42.80 (0.94)	43.26 (0.43)	45.65 [†] (0.11)
DMOZ-SMALL	μF_1 ($\uparrow$)	45.10 (0.23)	45.48 (0.19)	44.34 (0.32)	45.80 (0.64)	46.01 (0.74)	45.43 (0.21)	46.61 [†] (0.28)
	MF_1 ($\uparrow$)	30.65 (0.43)	30.60 (0.54)	30.94 (0.53)	30.62 (0.32)	30.82 (0.63)	30.34 (0.12)	31.86 [‡] (0.64)
DMOZ-2010	μF_1 ($\uparrow$)	40.22 (0.55)	41.32 (0.32)	40.34 (0.24)	41.77 (0.56)	41.82 (0.42)	40.71 (0.83)	42.37 (0.27)
	MF_1 ($\uparrow$)	28.37 (0.46)	29.05 (0.84)	28.41 (0.57)	29.11 (0.13)	29.18 (0.54)	28.66 (0.53)	30.41 (0.64)
DMOZ-2012	μF_1 ($\uparrow$)	50.13 (0.28)	50.32 (0.42)	50.11 (0.32)	48.05 (0.39)	50.31 (0.48)	49.90 (0.92)	50.64 (0.22)
	MF_1 ($\uparrow$)	29.89 (0.23)	29.89 (0.23)	29.73 (0.14)	27.65 (0.48)	30.04 (0.57)	30.52 (0.74)	30.58 (0.28)

Table shows mean and (standard deviation) in bracket across five runs. ‘X’ denotes MLF not possible. The significance test results are denoted as [†] for a p value less than 0.05 and [‡] for p value less than 0.01. We have used sign test and Wilcoxon rank test for statistical evaluation of μF_1 and MF_1 scores, respectively. Tests are between our best proposed approach, Global-INF and best baseline approach, MTA for single run. These statistical tests are not performed on DMOZ-2010 and DMOZ-2012 datasets because we do not have access to true labels from the online evaluation system

Table 5 μF_1 and MF_1 performance comparison of various top-down hierarchical baselines against our proposed approaches for SVM model

Dataset	Metric	Hierarchical baselines					Proposed approaches	
		TD-SVM	TLF	BLF	MLF	MTA	Level-INF	Global-INF
CLEF	μF_1 ($\uparrow$)	73.26 (0.13)	74.25 (0.16)	73.76 (0.36)	X	74.12 (0.32)	74.16 (0.58)	75.75 (0.01)
	MF_1 ($\uparrow$)	36.22 (0.01)	39.71 (0.27)	38.70 (0.14)	X	38.44 (0.36)	38.63 (0.42)	42.89 [‡] (0.01)
DIATOMS	μF_1 ($\uparrow$)	53.84 (0.42)	57.13 (0.42)	54.30 (0.31)	X	58.42 (0.12)	58.39 (0.43)	61.00 [‡] (0.39)
	MF_1 ($\uparrow$)	44.46 (0.24)	45.17 (0.62)	44.30 (0.64)	X	45.21 (0.65)	48.77 (0.12)	51.85 [‡] (0.23)
IPC	μF_1 ($\uparrow$)	50.63 (0.10)	51.36 (0.43)	50.55 (0.28)	X	51.38 (0.24)	50.43 (0.51)	52.48 [†] (0.11)
	MF_1 ($\uparrow$)	42.58 (0.12)	45.02 (0.42)	43.80 (0.26)	X	42.88 (0.52)	43.32 (0.62)	45.78 [†] (0.06)
DMOZ-SMALL	μF_1 ($\uparrow$)	46.12 (0.08)	46.63 (0.04)	46.29 (0.43)	46.58 (0.14)	46.20 (0.32)	46.43 (0.21)	47.39 [†] (0.16)
	MF_1 ($\uparrow$)	32.86 (0.11)	32.93 (0.28)	32.72 (0.44)	32.77 (0.52)	31.94 (0.28)	30.34 (0.12)	32.75 [‡] (0.34)
DMOZ-2010	μF_1 ($\uparrow$)	40.28 (0.28)	41.30 (0.38)	40.32 (0.18)	41.80 (0.32)	41.84 (0.10)	41.17 (0.24)	42.48 (0.23)
	MF_1 ($\uparrow$)	28.38 (0.15)	29.45 (0.62)	28.70 (0.36)	29.22 (0.04)	29.41 (0.26)	28.94 (0.53)	30.84 (0.42)
DMOZ-2012	μF_1 ($\uparrow$)	50.24 (0.29)	50.68 (0.37)	50.26 (0.11)	48.88 (0.21)	50.42 (0.28)	49.98 (0.12)	50.92 (0.24)
	MF_1 ($\uparrow$)	29.92 (0.14)	29.94 (0.35)	29.86 (0.23)	27.78 (0.57)	30.11 (0.51)	30.63 (0.18)	30.96 (0.12)

Table shows mean and (standard deviation) in bracket across five runs. ‘X’ denotes MLF not possible. The significance test results are denoted as [†] for a p value less than 0.05 and [‡] for p value less than 0.01. We have used sign test and Wilcoxon rank test for statistical evaluation of μF_1 and MF_1 scores, respectively. Tests are between our best proposed approach, Global-INF and best baseline approach, MTA for single run. These statistical tests are not performed on DMOZ-2010 and DMOZ-2012 datasets because we do not have access to true labels from the online evaluation system

Fig. 4 P value comparison of different approaches for LR model on DMOZ-SMALL dataset. Significant improvements are denoted by dark black color. **a** μF_1 . **b** MF_1

Table 6 Hierarchical performance comparison of various top-down hierarchical baselines against our proposed approaches over original (expert-defined) and modified hierarchy for LR model

Dataset	Hierarchy used	Hierarchical baselines				Proposed approaches		
		TD-LR	TLF	BLF	MLF	MTA	Level-INF	Global-INF
CLEF		hF ₁ score (↑)						
	Original	74.52 (0.01)	78.24 (0.75)	75.13 (0.46)	X	76.01 (0.74)	76.81 (0.59)	79.06 (0.01)
	Modified	–	77.78 (0.65)	78.08 (0.13)	X	77.50 (0.23)	78.28 (0.24)	80.87 (0.13)
	Original	56.15 (0.21)	62.53 (0.43)	56.14 (0.17)	X	59.60 (0.28)	60.03 (0.24)	62.80 (0.04)
DIATOMS	Modified	–	63.38 (0.24)	57.02 (0.62)	X	59.70 (0.14)	59.98 (0.28)	63.88 (0.13)
	Original	62.57 (0.32)	64.39 (0.38)	63.00 (0.10)	X	63.42 (0.54)	63.26 (0.34)	64.73 (0.12)
IPC	Modified	–	65.48 (0.32)	63.24 (0.41)	X	63.14 (0.54)	62.52 (0.38)	66.29 (0.28)
	Original	63.14 (0.54)	63.17 (0.43)	63.26 (0.52)	63.32 (0.64)	63.20 (0.54)	61.98 (0.56)	63.37 (0.44)
DMOZ-SMALL	Original	–	64.32 (0.50)	63.94 (0.38)	63.39 (0.19)	63.82 (0.42)	58.02 (0.14)	64.97 (0.75)
	Modified	73.04 (0.21)	72.70 (0.17)	73.04 (0.28)	70.49 (0.03)	73.03 (0.11)	71.41 (0.38)	73.19 (0.02)
DMOZ-2012	Original	TE score (↓)						
	Original	1.26 (0.01)	1.08 (0.08)	1.23 (0.03)	X	1.13 (0.09)	1.15 (0.05)	1.04 (0.03)
CLEF	Modified	–	0.89 (0.07)	0.88 (0.04)	X	0.90 (0.04)	0.94 (0.01)	0.71 (0.09)
	Original	1.76 (0.01)	1.49 (0.01)	1.76 (0.03)	X	1.60 (0.03)	1.60 (0.06)	1.49 (0.02)
DIATOMS	Modified	–	1.28 (0.02)	1.32 (0.02)	X	1.14 (0.06)	1.16 (0.02)	1.08 (0.08)
	Original	2.23 (0.02)	2.12 (0.04)	2.20 (0.01)	X	2.22 (0.06)	2.19 (0.01)	2.10 (0.02)
IPC	Modified	–	1.64 (0.01)	1.58 (0.04)	X	1.80 (0.06)	1.83 (0.03)	1.38 (0.02)
	Original	3.55 (0.04)	3.55 (0.02)	3.53 (0.03)	3.53 (0.06)	3.51 (0.08)	3.65 (0.06)	3.50 (0.02)
DMOZ-SMALL	Original	–	2.96 (0.05)	2.90 (0.01)	2.62 (0.03)	2.68 (0.03)	2.82 (0.02)	2.37 (0.03)
	Modified	3.69 (0.03)	3.58 (0.01)	3.68 (0.10)	3.56 (0.08)	3.61 (0.02)	3.74 (0.04)	3.53 (0.01)
DMOZ-2010	Original							
	Modified							

Table shows mean and (standard deviation) in bracket across five runs. 'X' denotes MLF not possible. Evaluations for DMOZ-2010 and DMOZ-2012 datasets cannot be performed on new hierarchy as it is not supported by the web portal. Further, hF₁ for DMOZ-2010 and TE score for DMOZ-2012 dataset is not available from the online evaluation system

Table 7 Hierarchical performance comparison of various top-down hierarchical baselines against our proposed approaches over original (expert-defined) and modified hierarchy for SVM model

Dataset	Hierarchy used	Hierarchical baselines				Proposed approaches		
		TD-SVM	TLF	BLF	MLF	MTA	Level-INF	Global-INF
CLEF		hF ₁ score (↑)						
	Original	74.91 (0.01)	76.23 (0.24)	75.33 (0.15)	X	76.18 (0.35)	76.22 (0.55)	77.62 (0.01)
	Modified	–	71.03 (0.28)	78.50 (0.48)	X	78.04 (0.14)	78.29 (0.08)	79.81 (0.06)
	Original	57.45 (0.13)	62.76 (0.22)	56.68 (0.26)	X	60.21 (0.84)	60.06 (0.14)	63.05 (0.01)
DIATOMS	Modified	–	63.46 (0.14)	58.24 (0.34)	X	60.74 (0.08)	60.10 (0.31)	63.94 (0.03)
	Original	62.86 (0.14)	64.46 (0.26)	63.24 (0.03)	X	63.47 (0.52)	63.58 (0.46)	65.14 (0.06)
IPC	Modified	–	65.50 (0.14)	63.32 (0.25)	X	63.86 (0.42)	63.18 (0.18)	66.42 (0.11)
	Original	62.86 (0.31)	64.34 (0.14)	62.83 (0.63)	64.36 (0.42)	64.61 (0.32)	64.66 (0.32)	65.50 (0.32)
DMOZ-SMALL	Modified	–	62.31 (0.35)	64.82 (0.09)	61.54 (0.04)	64.18 (0.42)	64.72 (0.51)	65.72 (0.20)
	Original	73.02 (0.45)	72.64 (0.24)	72.89 (0.17)	70.44 (0.09)	73.04 (0.02)	71.84 (0.26)	73.16 (0.06)
CLEF		TE score (↓)						
	Original	1.24 (0.01)	1.17 (0.03)	1.22 (0.01)	X	1.12 (0.11)	1.12 (0.04)	1.11 (0.02)
	Modified	–	0.86 (0.04)	0.85 (0.10)	X	0.88 (0.32)	0.85 (0.21)	0.73 (0.04)
	Original	1.74 (0.01)	1.45 (0.03)	1.74 (0.05)	X	1.58 (0.02)	1.53 (0.04)	1.48 (0.01)
DIATOMS	Modified	–	1.27 (0.03)	1.30 (0.02)	X	1.13 (0.04)	1.13 (0.06)	1.06 (0.01)
	Original	2.20 (0.04)	2.11 (0.03)	2.18 (0.02)	X	2.20 (0.02)	2.17 (0.03)	2.08 (0.01)
IPC	Modified	–	1.63 (0.02)	1.57 (0.03)	X	1.80 (0.04)	1.80 (0.02)	1.34 (0.06)
	Original	3.56 (0.22)	3.42 (0.05)	3.56 (0.11)	3.43 (0.02)	3.59 (0.08)	3.58 (0.14)	3.59 (0.10)
DMOZ-SMALL	Modified	–	2.75 (0.08)	2.86 (0.18)	2.30 (0.24)	2.66 (0.03)	2.70 (0.08)	2.39 (0.12)
	Original	3.67 (0.01)	3.54 (0.04)	3.60 (0.05)	3.52 (0.02)	3.54 (0.04)	3.56 (0.08)	3.49 (0.02)

Table shows mean and (standard deviation) in bracket across five runs. 'X' denotes MLF not possible. Evaluations for DMOZ-2010 and DMOZ-2012 datasets cannot be performed on new hierarchy as it is not supported by the web portal. Further, hF₁ for DMOZ-2010 and TE score for DMOZ-2012 dataset is not available from the online evaluation system

Fig. 5 MF_1 performance comparison of flat approach (marked in dotted line) against best top-down approach, Global-INF (marked in solid line) for LR model with different selection of threshold (τ) for CLEF and DMOZ-SMALL datasets. Validation data are used for plotting the graph

flattening of hierarchies results in the reduction of hierarchical path length for misclassified examples contributing to performance improvement.

6.3 Empirical study for threshold (τ) selection

Figure 5 shows the MF_1 performance comparison of flat LR approach against our best proposed approach (i.e., Global-INF) with varying selection of threshold (τ) in the interval $[\mu, \mu + 3\sigma]$ (performance deteriorates after $\mu + 3\sigma$) with step-size 0.1σ for CLEF and DMOZ-SMALL datasets. We choose these datasets for evaluation because they have different data characteristics. The CLEF dataset is well balanced and does not suffer from the rare categories issue, whereas DMOZ-SMALL dataset is highly imbalanced and majority of the classes belong to rare categories (i.e., having ≤ 10 examples) as shown in Fig. 2a. In order to identify the set of inconsistent nodes in the hierarchy, we compare the computed f_n^* value of each internal node with the chosen threshold (τ) and mark the node as inconsistent iff $f_n^* > \tau$. It can be seen from the figure that for CLEF dataset, performance improves as the threshold (τ) decreases giving intuition that τ should be kept smaller, i.e., removing more internal nodes from the hierarchy (enforcing flat structure) is better and hence reducing the threshold value τ can possibly lead to better results. However, for the DMOZ-SMALL dataset, performance first increases

and then decreases with maximum performance achieved at $\tau = \mu + 1.8\sigma$. This behavior suggests that for imbalanced data distribution with potentially large number of rare categories, we should generally keep the threshold higher. It helps to leverage the hierarchical information while reducing error propagation by removing inconsistent nodes. The best threshold for a specific dataset can be chosen empirically using a small validation set as done in this study.

To further understand the behavior of modified hierarchy using Global-INF approach, we analyzed the datasets in terms of level-wise fan-out (# children) in the hierarchy, before and after removing inconsistent nodes. We can see from Fig. 6 that with both datasets maximum flattening takes place at higher levels in the hierarchy, which results in increased fan-out value. Reason for inconsistencies at higher levels is the presence of many dissimilar classes beneath each node, which makes it comparatively difficult to learn generalized classifiers resulting in higher f_n^* values (i.e., inconsistent nodes marked for removal).

6.4 Level-wise misclassification error

Table 8 shows the level-wise error analysis that is obtained for TD-LR and our best approach, Global-INF for CLEF and DMOZ-SMALL datasets. We can see that at higher levels, the Global-INF approach misclassifies fewer examples (shown in #ME column) that results in less error propagation down the levels and hence better overall performance. This experiment supports our hypothesis that Global-INF approach identifies better set of inconsistent nodes that helps in minimizing the error propagation. Results for other top-down baselines are intermediate between these two and hence not shown in the paper for brevity.

6.5 Comparison to flat baselines

Table 9 shows the MF_1 and hF_1 performance comparison of our best approach, Global-INF against flat baselines—LR and ECOC approach. For easier analysis, we have showed the results for datasets separated by varying distribution of training size. (For evaluating DMOZ-2010 and DMOZ-2012

Fig. 6 Total fan-out (# children) at each level for CLEF and DMOZ-SMALL datasets before and after flattening inconsistent nodes using Global-INF approach. **a** CLEF. **b** DMOZ-SMALL

Table 8 Level-wise error for TD-LR and Global-INF approach

Dataset	Level no.	TD-LR		Global-INF	
		Error ($\downarrow$)	# ME	Error ($\downarrow$)	# ME
CLEF	L-1	21.27 (0.63)	214	20.18 (0.26)	203
	L-2	07.71 (0.42)	240	07.34 (0.29)	224
	L-3	11.30 (0.16)	274	05.66 (0.13)	227
DMOZ-SMALL	L-1	42.47 (0.32)	789	39.83 (0.17)	740
	L-2	14.45 (0.62)	921	12.91 (0.21)	855
	L-3	15.14 (0.34)	972	17.99 (0.14)	968
	L-4	12.32 (0.02)	1001	07.57 (0.10)	991
	L-5	15.66 (0.05)	1020	33.33 (0.04)	992

Table shows mean and (standard deviation) of error rate across five runs. # ME denotes the average number of misclassified examples up to that level

Table 9 MF_1 and hF_1 performance comparison between Global-INF and flat baselines with varying distribution of training examples per class

Dataset	# Train ex. per class	Best proposed		Flat baselines			
		Global-INF		LR		ECOC	
		MF_1	hF_1	MF_1	hF_1	MF_1	hF_1
CLEF	6–10	36.12	66.45	35.14	62.34	36.16	63.45
	11–50	45.52	76.13	45.92	77.04	45.36	76.40
	>50	52.24	84.12	55.92	87.24	53.24	85.90
	Avg.	46.99	79.00	51.31[‡]	80.58	50.02	79.34
DIATOMS	6–10	41.08	48.92	38.92	46.32	39.14	47.44
	11–50	42.82	62.72	44.18	62.45	43.82	59.26
	>50	53.17	64.10	57.24	68.10	52.21	63.28
	Avg.	51.85	62.80	54.17[‡]	63.50	48.82	61.92
IPC	11–50	43.64	62.45	43.94	60.92	42.98	60.01
	>50	47.28	68.70	49.44	68.95	47.21	67.96
	Avg.	45.65	65.73	45.74	64.00	44.65	61.84
DMOZ-SMALL	≤ 5	28.77	51.86	27.02	46.81	27.12	47.35
	6–10	55.55	67.47	54.76	65.40	54.18	63.28
	11–50	72.26	78.74	72.60	80.12	72.02	78.53
	>50	69.43	86.70	71.44	88.95	69.80	85.89
	Avg.	31.86[†]	63.37	30.80	60.87	30.10	60.50
DMOZ-2010	≤ 5	18.23	53.59	14.35	48.13	10.46	47.47
	6–10	23.03	55.76	22.62	51.84	20.74	49.63
	11–50	42.56	62.39	43.26	61.85	41.92	60.44
	>50	70.74	77.51	73.20	81.51	68.92	77.18
	Avg.	28.41[†]	56.17	27.06	53.94	26.12	49.24
DMOZ-2012	≤ 5	10.28	50.56	8.78	48.01	7.41	47.14
	6–10	20.37	50.71	18.84	48.82	18.32	48.26
	11–50	37.19	73.16	37.98	73.24	35.72	72.73
	>50	53.20	79.73	55.72	84.92	50.23	78.10
	Avg.	29.14[‡]	68.24	27.04	66.45	26.64	65.10

Bold values denote the best result across different methods

CLEF, DIATOMS and IPC datasets do not have any categories with ≤ 5 examples (≤ 10 for IPC). The significance test results are denoted as [†] for a p value less than 0.05 and [‡] for p value less than 0.01. Wilcoxon rank test is used for statistical evaluation of MF_1 scores. Tests are between Global-INF and best flat, LR approach

datasets, we have used a separate held-out dataset because we don't know the actual labels of test dataset from the online evaluation.) We show the results for MF_1 because it gives

equal importance to all the classes while evaluation and hence provides better essence of the results for datasets with skewed distribution. For computing hF_1 , we have used the original

Fig. 7 Performance improvement of hybrid models over baseline TD-LR model. For DMOZ-2010 dataset hF_1 score is not available from online web portal hence not reported. **a** MF_1 . **b** hF_1

hierarchy for consistent evaluation. Moreover, we perform significance test on the overall results (denoted by **avg.**) for each dataset. As we can see from the table, the LR approach outperforms Global-INF approach for CLEF, DIATOMS and IPC datasets because these datasets are well balanced and have smaller number of categories that needs to be discriminated which are easily separable by linear classifiers. However, for the DMOZ datasets, our approach Global-INF has better performance because hierarchical structure provides useful information for categorizing classes with rare categories. Within the DMOZ datasets, rare categories make up more than 75% of the classes as shown in Fig. 2. The ECOC approach has the worst performance because the code-words used in our experiments are chosen randomly and merging of categories may require nonlinear discriminants instead of the linear classifiers used in this paper.

6.6 Hybrid prediction

Figure 7 shows the hybrid prediction results for different LR models. It can be seen from the figure that hybrid method of prediction gives consistently better performance in comparison with the individual classifiers performance shown in Table 4. For example, CLEF dataset shows an improvement of $\sim 15\%$ in MF_1 score with hybrid model, TD-LR + LR as compared to the hierarchical model, TD-LR. Overall, Global-INF + LR model has the best performance for all datasets. Although hybrid models have better performance, one of the major drawbacks of hybrid approach is the expensive predic-

tion time (shown in Fig. 8) because it invokes all the models for leaf categories (for flat classifier prediction) as well as the relevant models corresponding to top-down prediction path (for hierarchical prediction).

6.7 Computational run time

Although the flat LR approach outperforms Global-INF approach for some datasets in terms of classification performance, their prediction runtime is significantly higher and it can be untenable for large-scale problems [3, 23]. The prediction runtime comparison of Global-INF and LR approach along with hybrid approach, Global-INF + LR, is shown in Fig. 8. As expected, Global-INF approach has comparatively lower prediction runtime (up to 4x improvement). The difference is significant for large-scale datasets (DMOZ-2010 and DMOZ-2012). For completeness, we also report the total training runtime in Table 10. The Global-INF approach has higher training runtime due to the overhead involved with classifiers retraining after hierarchy modification and also involves training one-vs-rest binary classifiers for internal nodes in addition to leaf categories. (It should be noted that for hybrid approach, Global-INF + LR, training runtime is same as Global-INF approach.) Nevertheless, both flat and top-down approaches are trivially parallelizable due to decoupling (i.e., no interactions) between the classifiers learnt at different nodes n in the hierarchy. For reporting training runtime, we trained classifiers in parallel across multiple compute nodes in the cluster and sum up the time

Fig. 8 Prediction runtime comparison (in mins) between TD, flat and hybrid approach. Image datasets have small difference hence omitted. **a** Small datasets. **b** Large datasets

Table 10 Total training runtime comparison (in mins) between Global-INF and LR approach

	CLEF	DIATOMS	IPC	DMOZ-SMALL	DMOZ-2010	DMOZ-2012
Global-INF	3	10	830	68	25,462	63,000
LR	1	3	658	46	15,248	46,124

taken at each node. In our experiments, we choose expensive one-vs-rest binary classifiers over comparably cheaper one-vs-sibling binary classifiers because our preliminary experiments showed better results with one-vs-rest approach. It should also be noted that there is no significant difference between the prediction and training runtime of different top-down approaches, and hence, we do not report them here.

7 Conclusion and future work

In this paper, we proposed two different approaches for hierarchy modification that restructures the hierarchy by flattening most prominent set of inconsistent nodes, thereby improving the hierarchy representation which is more suited for HC. Performance evaluation on wide range of datasets over the proposed modified hierarchy shows improved classification results because fewer examples are misclassified at higher levels, resulting in less error propagation. Comparison of our proposed approach with the competitive hierarchy modification approaches in the literature showed significant performance improvement supporting the hypothesis that our approach identifies the better set of inconsistent nodes. We also performed experiments to compare our approach with the flat approach with varying distribution of training examples per categories. Results demonstrated the usefulness of leveraging hierarchical information for classifying classes with fewer training examples. Further, we showed that combining the flat and top-down approach prediction results in an ensemble setting results in a much better performance but at the cost of expensive prediction time.

In future, we plan to study the effect of our modified hierarchy on various state-of-the-art HC approaches [3].

Acknowledgements The authors gratefully acknowledge the support of the work described in this paper from the NSF Grant #1447489 and #1252318.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Naik, A., Rangwala, H.: Inconsistent node flattening for improving top-down hierarchical classification. In: IEEE DSAA, pp. 379–388 (2016)
2. Silla Jr., C., Freitas, A.: A survey of hierarchical classification across different application domains. *Data Min. Knowl. Disc.* **22**(1–2), 31–72 (2011)
3. Gopal, S., Yang, Y.: Recursive regularization for large-scale classification with hierarchical and graphical dependencies. In: ACM SIGKDD, pp. 257–265 (2013)
4. Naik, A., Rangwala, H.: A ranking-based approach for hierarchical classification. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10 (2015)
5. Cai, L., Hofmann, T.: Hierarchical document categorization with support vector machines. In: CIKM, pp. 78–87 (2004)
6. Koller, D., Sahami, M.: Hierarchically classifying documents using very few words. In: ICML, pp. 170–178 (1997)
7. McCallum, A., Rosenfeld, R., Mitchell, T., Ng, A.: Improving text classification by shrinkage in a hierarchy of classes. In: ICML, pp. 359–367 (1998)
8. Naik, A., Rangwala, H.: Embedding feature selection for large-scale hierarchical classification. In: IEEE International Conference on Big Data, pp. 1212–1221 (2016)
9. Naik, A., Rangwala, H.: Large-scale hierarchical classification with rare categories and inconsistencies. *AI Matters* **2**(3), 27–29 (2016)
10. Dumais, S., Chen, H.: Hierarchical classification of web content. In: ACM SIGIR, pp. 256–263 (2000)
11. Sun, A., Lim, E.: Hierarchical text classification and evaluation. In: ICDM, pp. 521–528 (2001)
12. Xue, G., Xing, D., Yang, Q., Yu, Y.: Deep classification in large-scale text hierarchies. In: ACM SIGIR, pp. 619–626 (2008)
13. Xiao, L., Zhou, D., Wu, M.: Hierarchical classification via orthogonal transfer. In: ICML, pp. 801–808 (2011)
14. Zimek, A., Buchwald, F., Frank, E., Kramer, S.: A study of hierarchical and flat classification of proteins. *IEEE/ACM TCBB* **7**(3), 563–571 (2010)
15. Babbar, R., Partalas, I., Gaussier, E., Amini, M.: On flat versus hierarchical classification in large-scale taxonomies. In: NIPS, pp. 1824–1832 (2013)
16. Liu, T., Yang, Y., Wan, H., Zhou, Q., Gao, H., Zeng, B., Chen, Z., Ma, W.: An experimental study on large-scale web categorization. In: WWW Special Interest Tracks and Posters, pp. 1106–1107 (2005)
17. Wang, X., Lu, B.: Flatten hierarchies for large-scale hierarchical text categorization. In: ICDM, pp. 139–144 (2010)
18. Babbar, R., Partalas, I., Gaussier, E., Amini, M.R.: Maximum-margin framework for training data synchronization in large-scale hierarchical classification. In: Neural Information Processing, pp. 336–343 (2013)
19. Vens, C., Struyf, J., Schietgat, L., Džeroski, S., Blockeel, H.: Decision trees for hierarchical multi-label classification. *Mach. Learn.* **73**(2), 185–214 (2008)
20. Dekel, O.: Distribution-calibrated hierarchical classification. In: NIPS, pp. 450–458 (2009)
21. Tang, L., Zhang, J., Liu, H.: Acclimatizing taxonomic semantics for hierarchical content classification. In: ACM SIGKDD, pp. 384–393 (2006)
22. Dekel, O., Keshet, J., Singer, Y.: Large margin hierarchical classification. In: ICML, p. 27 (2004)
23. Liu, T., Yang, Y., Wan, H., Zeng, H., Chen, Z., Ma, W.: Support vector machines classification with a very large-scale taxonomy. *ACM SIGKDD Exp. Newslett.* **7**(1), 36–43 (2005)

24. Hadley, C., Jones, D.: A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure* **7**(9), 1099–1112 (1999)
25. Gao, T., Koller, D.: Discriminative learning of relaxed hierarchy for large-scale visual recognition. In: *ICCV*, pp. 2072–2079 (2011)
26. Naik, A., Rangwala, H.: Filter based taxonomy modification for improving hierarchical classification. [arxiv: 1603.00772](https://arxiv.org/abs/1603.00772) (2016)
27. Eisner, R., Poulin, B., Szafron, D., Lu, P., Greiner, R.: Improving protein function prediction using the hierarchical structure of the gene ontology. In: *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 1–10 (2005)
28. Charuvaka, A., Rangwala, H.: Hiercost: improving large scale HC with cost sensitive learning. In: *ECML PKDD*, pp. 675–690 (2015)
29. Naik, A., Charuvaka, A., Rangwala, H.: Classifying documents within multiple hierarchical datasets using multi-task learning. In: *ICTAI*, pp. 390–397 (2013)
30. Naik, A.: Using multi-task learning for large-scale document classification. Master's thesis, George Mason University, Fairfax, Virginia (USA) (2013)
31. Hernandez, J.N., Sucar, L.E., Morales, E.F.: A hybrid global-local approach for hierarchical classification. In: *FLAIRS Conference* (2013)
32. Dimitrovski, I., Kocev, D., Loskovska, S., Džeroski, S.: Hierarchical annotation of medical images. *Pattern Recogn.* **44**(10), 2436–2449 (2011)
33. Dimitrovski, I., Kocev, D., Loskovska, S., Džeroski, S.: Hierarchical classification of diatom images using predictive clustering trees. *Ecol. Inf.* **7**, 19–29 (2012)
34. Yang, Y.: An evaluation of statistical approaches to text categorization. *Inf. Retr.* **1**(1–2), 69–90 (1999)
35. Kosmopoulos, A., Partalas, I., Gaussier, E., Paliouras, G., Androutsopoulos, I.: Evaluation measures for hierarchical classification: a unified view and novel approaches. *Data Min. Knowl. Disc.* **29**(3), 820–865 (2015)
36. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: LIBLINEAR: a library for large linear classification. *JMLR* **9**, 1871–1874 (2008)
37. Ghani, R.: Using error-correcting codes for text classification. In: *ICML*, pp. 303–310 (2000)
38. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: *ACM SIGIR*, pp. 42–49 (1999)